Journal of Advances in Computer Research
Quarterly pISSN: 2345-606x eISSN: 2345-6078
Sari Branch, Islamic Azad University, Sari, I.R.Iran
(Vol. 6, No. 4, November 2015), Pages: 125-137
www.jacr.iausari.ac.ir

Proposing a Method to Classify Texts Using Data Mining

Mohammad Rostami✉[1], Seyed Saeed Ayat[2], Iman Attarzadeh[3], Farid Saghari[4]

1) Member of Young Researchers Club, Islamic Azad University, Dehaghan Branch, Isfahan, Iran
2) Associate Professor, Department of Computer Engineering and Information Technology, Payame Noor University
3) Department of Computer Engineering, Islamic Azad University, Dezfoul Branch, Iran
4) MS c, software engineering department

mohammadrostami@dehaghan.ac.ir; dr.ayat@pnu.ac.ir; attarzadeh.std@yahoo.com; 7lac.net@gmail.com

Received: 2015/03/06; Accepted: 2015/06/10

Abstract

Today a significant part of available data is saved in text database or text documents. The most important thing is to organize these documents. One way to organize text documents is to classify them. To classify texts is to assign text documents to their actual categories. This has two main steps, i.e. feature- and learning algorithm selection. There have been several methods suggested to classify text documents. In this paper, we propose a combined method to do this more efficiently. When selecting features, the proposed method uses filtering in order to reduce complexity and it is implemented using naïve Bayes and decision tree categories. Results indicate advantages of this combined method to individual classifying.

Keywords: *data mining, text documents, classify, decision tree, Bayes.*

1. Introduction

We live in a world, where information has a lot of value for us. Having increased the amount of information available online, the need for tools which can help us to search, filter and manage the resources is quite evident.

Text classification is thematic labeling of texts in natural languages based on a preset complex. Today, text classification is used in many fields, from text indexing based on a controlled dictionary, to text filtering, automatic metadata production, word sense disambiguation, production of hierarchical catalogues from web resources and in general in any application requiring document organization or a specific selective and comparative distribution of documents [1]. Other applications of text classification are automatic answering systems, information filtering, theme distinguish of data, spam emails, title recognition and other related fields [2].

Data are classified through classification methods based on their specifications. This pattern can be used to understand existing data and predict their behavior [20]. Automated text document classification is one of the most important methods to organize information and discover the knowledge hidden in the data as access to electronic documents and internet use have been expanding at a very high rate. Proper classification of text documents, online news, email and digital libraries require text searching, machine learning and natural language processing techniques to acquire

meaningful knowledge. Text searching techniques have recently grown in importance since access to electronic documents through different sources has increased. Unstructured and semi structured sources include the internet, news stories, biological databases, chat rooms, digital libraries, online forums and email. Therefore, correct classification and discovery of knowledge from such sources are among the most important issues. A combination of natural language processing, data mining and machine learning is used to classify and discover patterns from text documents [7]. The main objective of text searching is to extract information from text sources by means of operations such as similarity, retrieval, classification (with or without a supervisor or semi supervised) and summarization. Thus, one must be able to correctly classify and organize such documents. Therefore, many challenges such as proper annotation, proper document display and dimension reduction are confronted [5].

Today, a good part of existing information is saved in text databases (or text documents) which are composed of a large set of different documents and sources such as news, scientific papers, books, digital libraries, email messages and web pages. Text searching is the knowledge to extract information from unstructured text. In other words, one could consider text searching as methods and algorithms of machine learning fields and statistical techniques aiming at finding useful patterns in text.

One of the most important techniques of text searching is automated text classification [17]. Automated text classification means attributing existing text documents to some predefined classes to which the documents belong. To do so, first, the classes must be identified which is usually done by experts. Then, the text documents belonging to each class have to be determined. The main objective is to find the true classes related to collected documents. As internet and electronic text documents have become popular, automated text classification has turned into a must [4], [8], [19].

The main challenge of document classification is enlargement of features space in these issues. This large space in most available algorithms causes the classifier to be very slow and inefficient. In addition, there exist some features that not only lead not to better classification of documents but also reduce its accuracy [3].

A text cannot be interpreted directly by a classifier or a classifier algorithm; but it is mapped using an indexing process which modifies the text to an array (that its content will be given by the size). This helps providing integrity and unification of training set, experimental and validation texts [1].

The large size of term spaces in text classification is usually trouble some. In fact, increasing the number of terms leads to increase in numbers of features that causes more complexity (more time consumption and more memory space); on the other hand independence of data related to each other (information becomes less) generally doesn't worth classification.

In other words, it is tried in decreasing general size to identify terms with less value, analyzing language corpus and texts in training set; then all these terms will be determined in a fixed list. Terms of input text are filtered automatically by this list. In local area, this is done for each group separately. Therefore, size reduction is itself a good field in information recovery and in particular in classification [2].

The first approach of size reduction using term selection is called filtering approach. Using some tools provided by information or statistics theory, irrelative terms would be filtered from extracted terms. Finally, independent of the applied filtering function, classifiers will be produced using reduced term space [2].

A combination method using the decision tree and simple bees classifications has been offered in this article. As shown by the results, combination classifications can be more efficient compared to individual ones.

Assign the text documents to pre-define categories called Automatic text categorization. This is important which this work to be with high performance. In this paper to improve efficiency of text categorization we proposed a hybrid method which uses from a filtering method to reduce the complexity of feature selection and uses from J48 in the learning step. The proposed method uses same classifiers with different sampling with replacement from the training set. We compared proposed method with J48 and Naive Bayes single classifiers. The results show that the proposed method has better performance than single classifiers in average precision, average recall and average F1 criteria.

The remaining of the paper is organized as follows: in section 2 and 3 and 4 discusses the process of document classification. section 3 includes the proposed method and details of implementation and analysis of the proposed method. section 4 and 5 are dedicated to conclusion and references respectively.

2. A Review on Text Classification

According to the large amount of electronic textual information that are significantly available through internet and other sources, lacking appropriate indexing and classification causes information processing and recovery to face a lot of problems. Text classification has different applications including document pursuit, document management, document expand and information reduction. Several machine learning methods regarding text classification are used in recent years e.g. regression models [10], K-nearest neighbor (K-NN) [11], Naïve Bayes net works [12] and decision tree [13], each has a different calculation and accuracy.

In [14] classification of the texts in Turkish are examined using n-gram. In this research, texts are classified using bigram, unigram, trigram and quadgram. Tests are performed on six hundred text documents which are classified in six categories and the efficiency of this study is reported 85.83%.

Text classification using a combined algorithm is proposed in [15]; this algorithm is a combination of the algorithms K-NN and SVM. Results of this research done on a data set of Reuters indicate that the efficiency of this method is 81.48% in the best and 54.55% in the worst condition.

3. Types of Classification

According to classes discussed here, classification is divided to binary classification and multi-class classification. Binary classification allocates the samples exactly to one of the two available classification classes (Figure 1), while multi-class classification deals with more than two classes (Figure 1). Note that in this paper multi-class classification is discussed.

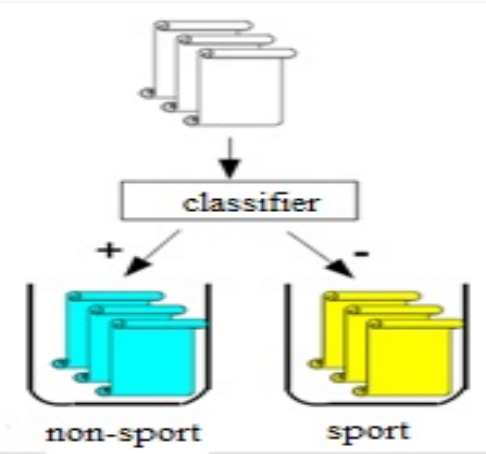

Figure 1. Binary Classification

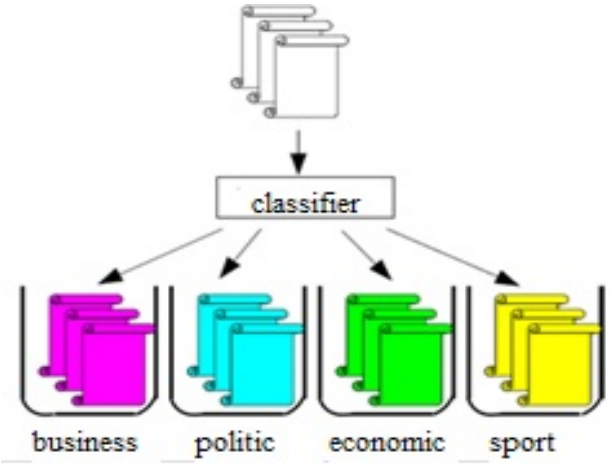

Figure 2. Multi-class Classification

4. Document Classification Process

Figure 3 shows the different stages of text classification including the set of documents, preprocessing, indexing characteristics, filtering characteristics, learning algorithm and evaluation [4], [8], [17], [19].

4.1 Documents

The first step includes collecting data of different formats such as pdf, doc, html.

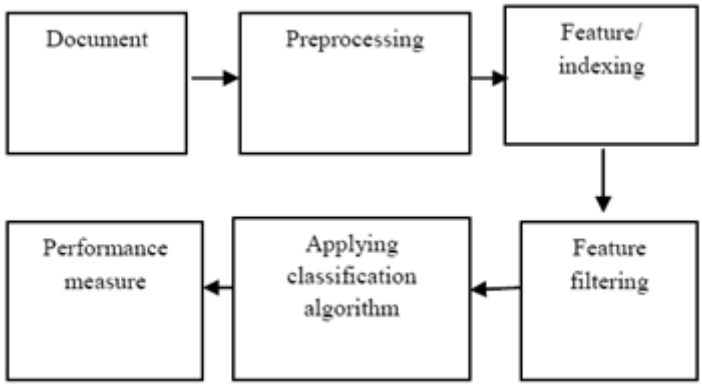

Figure 3. Different stages of automated text classification

4.2 Preprocessing

Data mining is the process of extracting hidden patterns from large data sets. Data in the real world are usually incomplete and incompatible with high error probability. At this stage, different things, such as the following are done to data in order to refine them [4], [8], [17], [18], [19].

- Tokenization: The whole text is changed into a set of distinguishable words leaving out punctuation and tags.
- Stop word removal: Words such as the, and, not, for (in English) are omitted at this stage.
- Term stemming: Words are changed into their root forms, and words that are differentiated with prefixes and suffixes but have the same root are placed in the same group (called words of the same family).

4.3 Term Stemming

At this stage of the process, a document is changed from text to document vector. One of the most common methods to do so is called the vector space model where documents are shown as vectors of words. After the above stages, the document can be shown as a vector of words with text characteristics. These characteristics can now be weighted according to their importance compared to their text document and class. The more important the characteristics, the more weighted.

4.4 Characteristic Selection

This stage is dedicated to selecting a subset of text characteristics. After the above mentioned stages, the number of characteristics in the text will be very large which will greatly reduce classification efficiency; therefore, the characteristic selection tries to pick the most important and most essential of the existing characteristics. Classification efficiency can improve by omitting unrelated and undistinguishable characteristics.

4.5 Learning Algorithm

Upon completion of the above mentioned stages, the text is now a vector of characteristics with different weights which is given to the learning algorithm to

produce a classification model. Upon production of the classification model, the class related to new texts can be distinguished.

4.6 Evaluation

The efficiency of classification can be determined using some parameters. The most important parameters include: correctness, accuracy, convocation and F criterion. Figure 4 shows steps method and research objective.

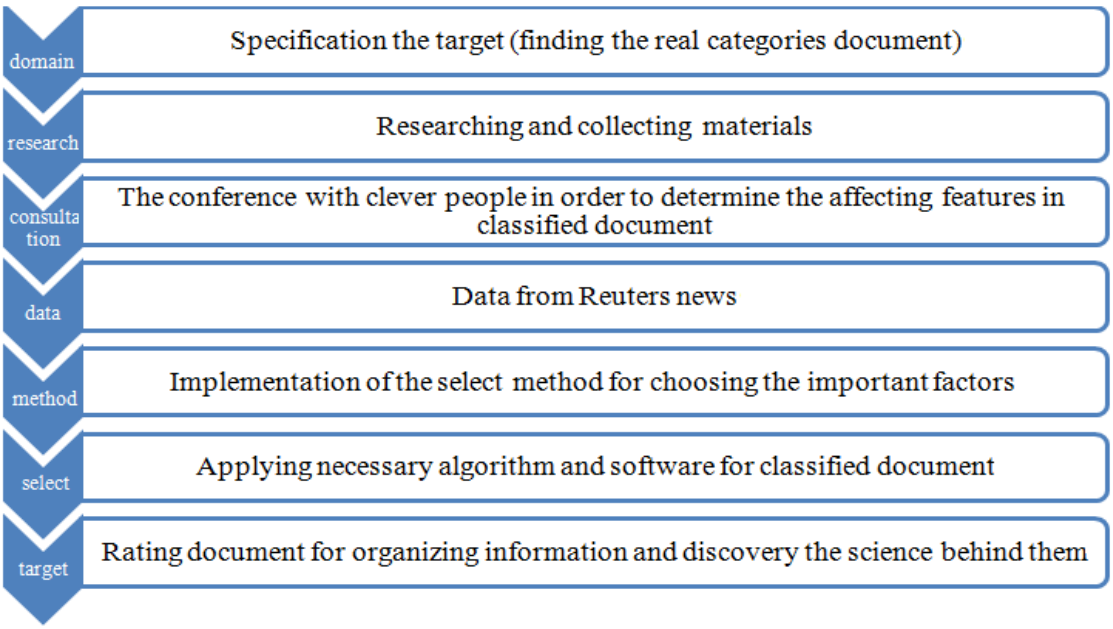

Figure 4. steps method and research objective

5. Proposed Model

A combination method using two individual basic classifiers is implemented and evaluated in this article considering the importance of automated text classification. The main idea is that a multi-info decision (classifier) can be better than a single-info decision. May of the less important characteristics are omitted in the preprocessing stage. Since characteristic selection methods are divided into two classes of filter and cover, the proposed method uses the information interest filter characteristic selection technique. Filter methods are less complicated compared to cover methods. For the learning stage, simple bees algorithms and a J48 decision tree is used. In the combination method proposed, classes are trained using different samplings, and a final decision is made to determine document class using the voting combination method in the end. Then, the efficiency of the proposed method is evaluated by using various evaluation criteria. The details of the proposed method are given below. Figure 5 shows diagrams research general.

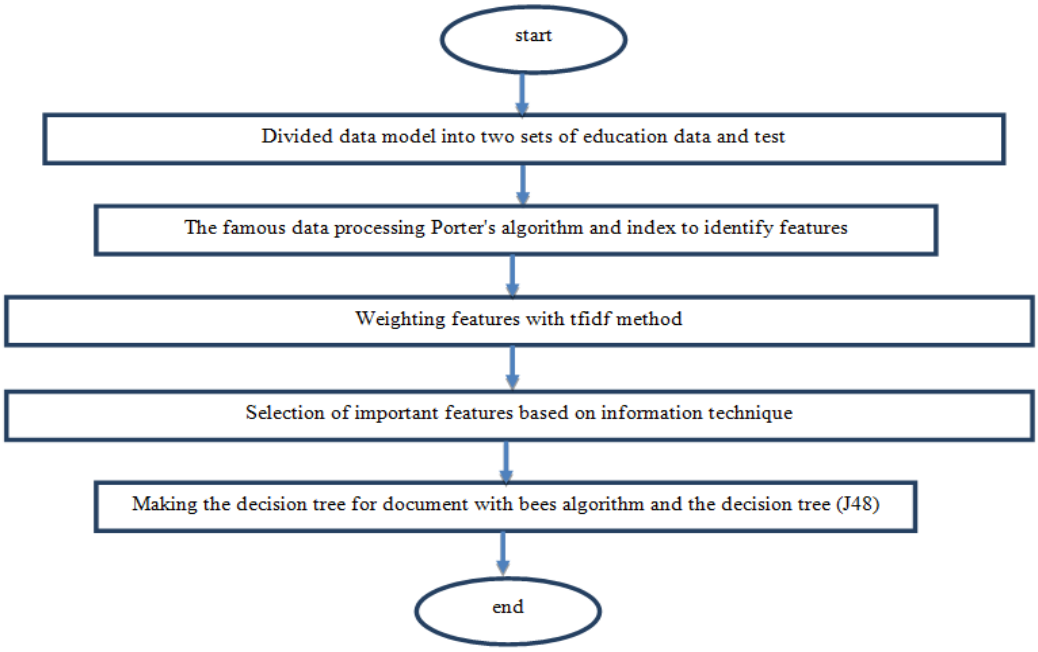

Figure 5. diagrams research general

5.1 Data Set

The real data set of Reuter's news – 21578 has been selected in this article. This set of data was collected in and indexed by a group of Reuter's News Agency in 1987. The original version of this data set includes 21578 text documents including Reuter's news with 118 different classes. The ModApte Split method is used to separate training and experimental documents [16]. Since the single-label – multi-class method is focused on in this article, an R(8) subset is used which includes 8 main classes and 7676 text documents. Each document belongs only to one of the existing classes [6]. Table 1 shows list of data set from Reuters for examine and test phase.

Table 1. Document List of the Examine and Test Phase for the Data Set Reuters-21578

Sets	Examine phase	Test Phase	Total
Acq	1596	696	2292
Trade	251	75	326
Ship	108	36	144
Interest	190	81	271
Grain	41	10	51
Crude	253	121	374
Earn	2841	1083	3924
Money-f_x	206	87	293

5.2 Preprocessing

In the preprocessing stage, operations such as changing capital letters to lower case in order to harmonize words, word separation, removing extra words, term stemming by well-know Porter's algorithm and n-gram indexing at 2 to change text to a set of words measuring 2 in length have been carried out.

5.3 Characteristic Weighting

The tfidf weighting method is used in this article where the functioning of classifiers is evaluated using this weighting method [9]. Tfidf is the result of multiplying two factors: tf and idf. Tf considers the repetition of the word in text while idf is document's reverse frequency. Equation 1 shows how this method is used to calculate.

$$Tfidf=tf*\log(N/n_i) \quad (1)$$

5.4 Characteristic Selection

The information interest technique has been used to select important characteristics. The usefulness of a word in a document in this method is the number of information bits calculated to predict class based on the existence or non-existence of the word in the text document [21]. Equation 2 shows this method.

$$IG(t_k,c_i)=\sum_{c\in\{c_i,\bar{c}_i\}} \sum_{c\in\{t_k,\bar{t}_k\}} p(\bar{t},c)\log_2\frac{p(t,c)}{p(\bar{t})p(c)} \quad (2)$$

Where p(t,ci) is the number of text documents within the class ci with the word (or letter) t in them, and p(t-,ci) indicates the number of text documents within the class ci without the word t.

5.5 Learning Algorithm

Various algorithms such as bees classification, decidion tree, closest proximity k, backup vector machine, neural networks, and the Rocchio algorithm have been proposed. This article uses the simple bees method and J48 decision tree. Another algorithm proposed in classification is the simple bees method. This method is important for different reasons. It has a very simple structure and requires no complicated repeated parameter estimation. That means it can be used for very large data sets. The bees method is one of the fastest algorithms in classification. It is based on conditional probabilities. The decision tree is one of the most well-known and most widely used methods in classification. In the classification model based on the decision tree, the output knowledge is offered as a tree of different modes of characteristic values. Depicting knowledge as a tree has made it possible for classifications based on the decision tree to be fully explainable. The classifier in the decision tree is a tree where internal nodes represent characteristics, edges protruding from nodes are the characteristic selection criteria, and leaves represent classes.

5.6 Classification Evaluation

For evaluation purposes, the label attributed by the classification model to the text document has to be compared to the label to which the text document belongs. The occurrence different modes for classes and documents are based on the input data sets for classification at FP, TP, FN, TN for the positive and negative classes. Different evaluation criteria have been offered some of the most important of which are the correctness criterion, accuracy, convocation and F1 criterion shown respectively in Equations 3 to 6 [19].

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+FN+TN} \quad (3)$$

$$\text{Precision}_i = \frac{TP_i}{TP_i + FP_i} \quad (4)$$

$$\text{Recall}_i = \frac{TP_i}{TP_i + FN_i} \quad (5)$$

$$\text{F1}_i = \frac{2 * Precision_i * Recall_i}{precision_i + Recall_i} \quad (6)$$

In the above equations, TP represents the number of samples with positive real class which have correctly been attributed to the positive class by the classification algorithm. FT represents the number of samples with negative real class which have wrongly been attributed to the positive class by the classification algorithm. TN represents the number of samples with negative real class which have correctly been attributed to the negative class by the classification algorithm. FN represents the number of samples with positive real class which have wrongly been attributed to the negative class by the classification algorithm.

This article uses the data mining software RapidMiner ver. 5.2 to implement the proposed method [22]. It is an open source data mining software written in Java language and developed since 2001 [20]. Figure 6 shows classification correctness average; Figure 7 shows classification accuracy average; Figure 8 convocation average and Figure 9 F1 criterion average. As can be seen, combination classifications are more efficient compared to individual classifications. Among those, the decision tree classifier functions better than the simple bees method.

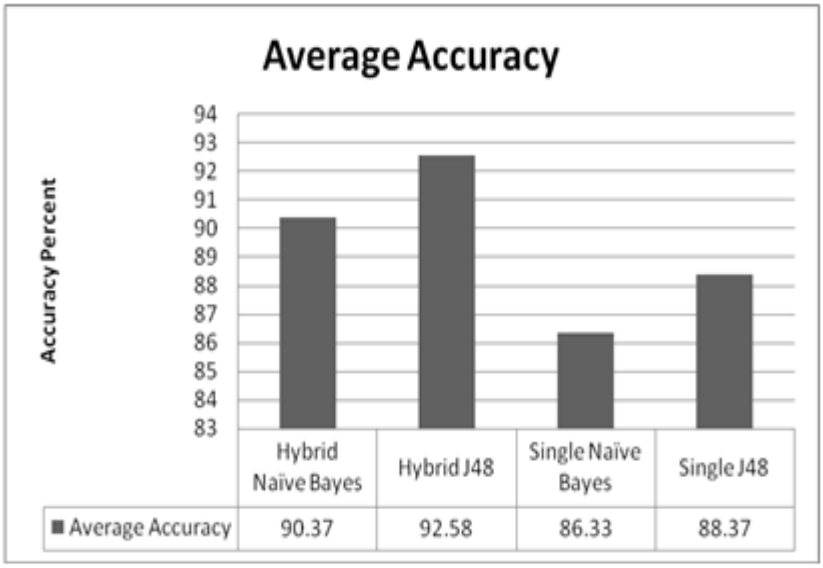

Figure 6. Classification correctness average

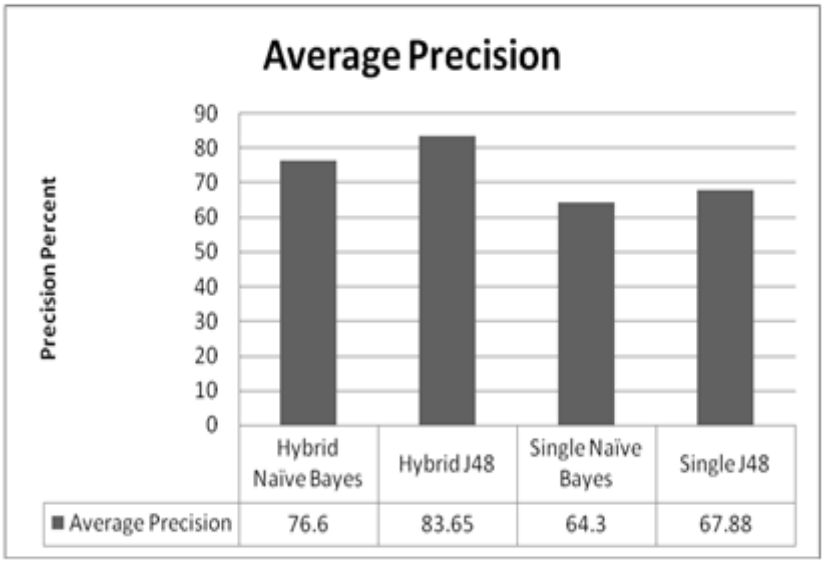

Figure 7. Classification accuracy average

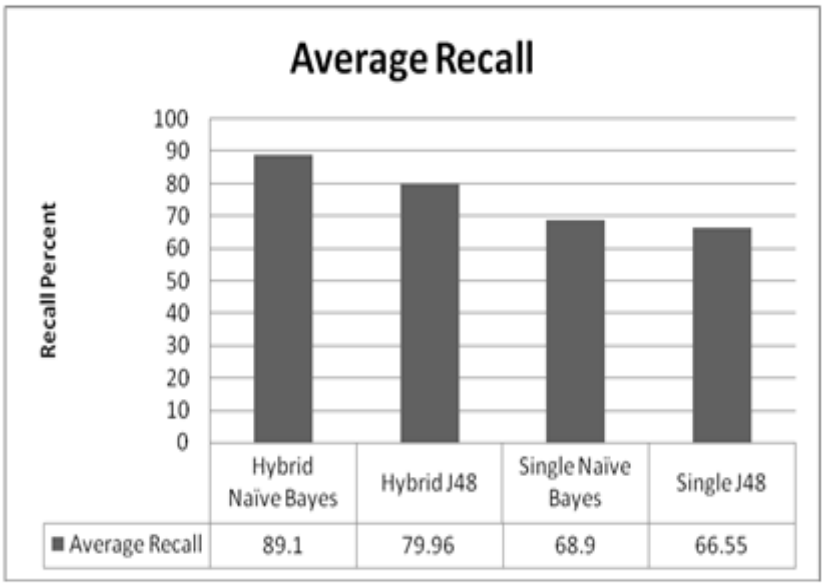

Figure 8. Average convocation

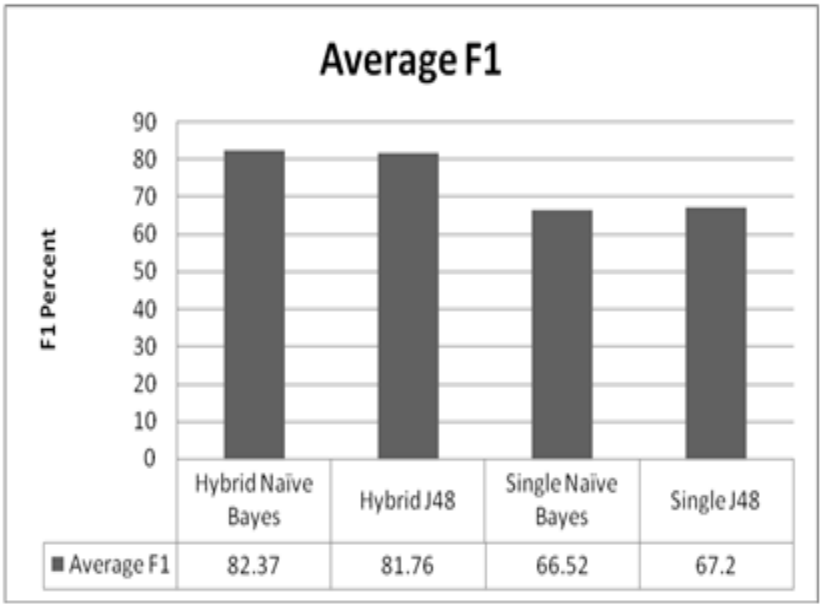

Figure 9. F1 criterion average

Results indicated that for this data set, the proposed method has the best efficiency to other algorithms. results shown in table 2 indicate that the best efficiency of the proposed method.

Table 2. Results of the Proposed Algorithm

	Average accuracy	Average precision	Average recall	Average F1
Hybrid Naive Bayes	90.37	76.60	89.10	82.37
Hybrid J48	92.58	83.65	79.96	81.76
Single Naive Bayes	86.33	64.30	68.90	66.52
Single J48	88.37	67.88	66.55	67.20

Table 3 shows the results of proposed method and J48 and Naive Bayes classifiers.

Table 3. results of proposed method and J48 and Naive Bayes classifiers

	Proposed Method	J48	Naive Bayes
Average Precision	82.55	66.88	64.30
Average Recall	78.96	66.56	70.10
Average F1	81.74	68.20	66.52

The results show that proposed method have better performance than J48 and Naive Bayes classifiers with %82.55 for average precision, %78.96 for average recall and %81.74 for average F1. Also for single classifiers, J48 have better performance than Naive Bayes classifier with %66.88 for average precision and %68.20 for average F1 but Naive Bayes has better performance than J48 with %70.10 for average recall. According to this note that average F1 calculated by average precision and average recall, we see that better performance related to the proposed method, J48 and finally Naive Bayes respectively.

6. Conclusion

In this paper, a new method of text classification is proposed. With growing access to the internet and text documents, it seems necessary to organize documents. One way to organize text data is to classify them. The objective of classifying text documents is to find the real class of the text document. A combination method using the decision tree and simple bees classifications has been offered in this article. As shown by the results, combination classifications can be more efficient compared to individual ones. Also, the decision tree classifier has shown better functioning than the simple bees method.

7. References

[1] S. Eyheramendy, A. Genkin, W. H. Ju, D. D. Lewis, D. Madigan, ”Sparse Bayesian Classifiers for Text Categorization”, Joint Statistical Meeting in San Francisco, California, 2003.

[2] I .Guyon, A. Elisseeff, ”An Introduction to Variable and Feature Selection”, Journal of Machine Learning Research,Vol. 3, 1157-1182, 2003.

[3] Y .Lin, Y. Qu, Z. Wang, ”A Novel Feature Selection Algorithm for Text Categorization”, Expert Systems with Applications, Vol. 33, 1–5, 2007.

[4] M.K. Dalal, and M.A. Zaveri, "Automatic Text Classification: A Technical Review", International Journal of Computer Applications, 28 (2), 37-40, 2011.

[5] A. Dasgupta, "Feature selection methods for text classification", In Proceedings of the 13th ACMSIGKDD international conference on Knowledge discovery and data mining, pp, 230-239, 2007.

[6] Dataset available in: http://csmining.org/index.php/r52-and-r8-of-reuters-21578.html, Retrieved at 2013/06/04.

[7] A. Khan, B. Baharudin, L.H. Lee, and K. Khan, "A Review of Machine Learning Algorithms for Text-Documents Classification", Journal of Advances in Information Technology, 1(1), 4-20, 2010.

[8] V. Korde, and C.N. Mahender, "TEXT CLASSIFICATION AND CLASSIFIERS: A SURVEY", International Journal of Artificial Intelligence & Applications (IJAIA), 3 (2), 85-99, 2012.

[9] M. Lan, and C.L. Tan, "Supervised and Traditional Term Weighting Methods for Automatic Text Categorization", Journal of IEEE Pami, 10 (10), 1-36, 2007.

[10] http://www.rapidi.com.

[11] C.H. Wan, L.H. Lee, R. Rajkumar, D. Isa,” A Hybrid Text Classification Approach with Low Dependency on Parameter by Integrating K-nearest neighbor and Support Vector Machine”, Elsevir, 2012.

[12] J. Sreemathy, P.S. Balamurugan,” An Efficient Text Classification Using KNN and Naïve Bayesian”, International Journal on Computer Science and Engineering (IJCSE), Vol. 4 No. 03, March 2012.

[13] Y.H. Li and A.K. Jain, “Classification of text documents”.The Computer Journal 41(8), 537-546, 1998.

[14] A. Guran, S. Akyokus, N.G. Bayazit, M. Zahidbgurbuz, ”Turkish Text Categorization Using n-gram word”, International Symposium on Innovations in Intelligent Systems and Applicaitons, June 29 – July 1, 2009.

[15] C.H. Wan, et al. “A hybrid text classification approach with low dependency on parameter by integrating K-nearest neighbor and support vector machine”. Expert Systems with Applications (2012), doi:10.1016/j.eswa.2012.02.068. Elsevir 2012.

[16] C.D. Manning, P. Raghavan and H. Schutze, "Introduction to information retrieval", Cambridge University Press, Cambridge, England, 2009.

[17] A. Patra, and D. Singh, "A Survey Report on Text Classification with Different Term Weighing Methods and Comparison between Classification Algorithms", International Journal of Computer Applications, Vol. 75, No.7, 14-18, 2013.

[18] S. Ramasundram," text categorization by Back propagation", International Journal of Computer Applications, Vol. 8, No. 6, 1-5, 2010.

[19] F. Sebastiani, "Machine Learning in Automated Text Categorization", ACM Computing Surveys, 34 (1), 1–47, 2002.

[20] L. Wang, and X. Fu, "Dtat Mining with Computational Intelligence", Advanced Information and Knowledge Processing, Springer-Verlag Berlin Heidelberg, 2005.

[21] Y. Yang, and J.A. Pedersen, "A comparative study on feature selection in text categorization In Proceedings of the 14th International Conference on Machine Learning (ICML-97)", 412-420, 1997.

[22] Software available in: http://rapid-i.com/content/view/181/190/lang,en/, Retrieved at 2013/06/04.